Desalination 260 (2010) 129–136

Contents lists available at ScienceDirect

Desalination

journal homepage: www.elsevier.com/locate/desal

Multivariate statistical analysis of surface water quality based on correlations and variations in the data set

R. Noori [a,b,*], M.S. Sabahi [b], A.R. Karbassi [b], A. Baghvand [b], H. Taati Zadeh [b]

[a] Department of Water Resources Research, Water Research Institute, Ministry of Energy, P.O. Box: 16765-313, Tehran, Iran
[b] Department of Environmental Engineering, Graduate Faculty of Environment, University of Tehran, P.O. Box 14155-6135, Tehran, Iran

ARTICLE INFO

Article history:
Received 8 February 2010
Received in revised form 19 April 2010
Accepted 24 April 2010
Available online 31 May 2010

Keywords:
Principal component analysis
Canonical correlation analysis
Karoon River
Water quality

ABSTRACT

In the research, determination of principal and non-principal monitoring stations was carried out using principal component analysis (PCA) technique for the Karoon River, Iran. Also canonical correlation analysis (CCA) was used to determine relationship between physical and chemical water quality parameters. Water quality parameters including BOD_5, COD, EC,NO_3^-, SO_4^{-2}, temperature, Cl^-, DO, hardness, TDS, pH, and turbidity were measured in samples collected from 17 stations along Karoon River from 1999 to 2002. Four of our monitoring stations proved less telling in explaining the annual variation of the river water quality, and were removed. Further investigations indicated that all water quality parameters were important. In CCA, the first four canonical correlations were 0.993, 0.822, 0.785, and 0.660, respectively, suggesting that EC and TDS were two dominant physical parameters in the all canonical variates whilst ions and hardness were highly scored from chemical parameters. Verifying the ability of PCA and CCA methods was carried out by simple regression and correlation methods, respectively.

1. Introduction

Surface water pollution by chemical, physical, and biological contaminants all over the world can be considered as an epidemic problem [1,2]. Surface water systems are waters naturally open to the atmosphere, such as rivers, lakes, reservoirs, estuaries, and coastal waters. Quality of a river at any point reflects several major influences, including the lithology of the basin, atmospheric inputs, climatic conditions and anthropogenic inputs [3]. Besides, rivers play a major role in assimilation or transporting municipal and industrial wastewater and runoff from agricultural lands [4]. Therefore, river water quality assessment is of great importance because it directly influences public health (via drinking water) and aquatic life (via raw water). During the last decades, river water quality monitoring by measurement of various water quality parameters has been increasing. However, due to spatial and temporal variations in water quality, which are often difficult to interpret, a monitoring program providing a representative and reliable estimation of the quality of surface waters, is necessary [5].

Creating a water quality monitoring system with appropriate efficiency is one difficulty in evaluating surface water quality; and to measure variables that would express water quality changes as much as possible. To achieve this goal, a multivariate statistical method such as principal component analysis (PCA) can be utilized. Recently, PCA has been widely used to evaluate a variety of environmental issues. Facchinelli et al. [6] used multivariate statistic methods such as PCA and cluster analysis (CA) to predict potential non-point heavy metals sources in soil on the regional scale. For the presentation of data, GIS supported software was used. Coming up with effective pollution control management for the surface water, Simeonov et al. [7] using PCA, clustering analysis (CA) and principal component regression interpreted a large and complex data matrix of surface water parameters in Northern Greece. Gangopadhyay et al. [8] applied the PCA and principal factor analysis (PFA) techniques to identify importance of monitoring wells predicting the dynamic variations related to potentiometric head at a location in Bangkok, Thailand. Through the PCA and GIS approaches, Terrado et al. [9] analyzed the main contamination sources of heavy metals, organic compounds, and other physicochemical parameters in Ebro River surface waters. Ouyang [10] adopted PCA and PFA to identify important water quality

* Corresponding author. Department of Water Resources Research, Water Research Institute, Ministry of Energy, P.O. Box: 16765-313, Tehran, Iran. Tel.: +98 2177000305; fax: +98 2177000910.
E-mail address: roohollahnoori@gmail.com (R. Noori).

0011-9164/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.desal.2010.04.053

parameters in twenty-two stations, located at the main stem of the lower St. Johns River in Florida, USA. Results revealed that total organic carbon, dissolved organic carbon, total nitrogen, dissolved nitrate and nitrite, orthophosphate, alkalinity, salinity, Mg, and Ca were the most important parameters in assessing variations of water quality in the river. Sherestha and Kazama [4] applied CA, PCA, PFA, and discriminant analysis techniques to evaluate temporal and spatial variations of a large complex water quality data set of the Fuji River basin.

Additionally, exploring relationships between physical and chemical parameters of water can help in river water quality management. Turbidity (Turb), total suspended solids (TSS), total dissolved solids (TDS), and so on originate from physical characteristics of rivers, whilst chemical oxygen demand (COD), sulfate (SO_4^{-2}) and nitrate (NO_3^-) ions, usually indicate water pollution by human activity. The study explores the relationship between physical and chemical parameters using canonical correlation analysis (CCA). CCA was utilized to ascertain the extent to which one set of measurements was related to another and to determine the particular attributes responsible for the relationships. The method is an extension of PCA technique [11]. Glahn [12] was the first to use CCA technique to forecast air temperature over United States from sea surface temperature and sea level pressure. Statheropoulos et al. [13] applied CCA technique to determine relationship between both data sets i.e. air pollutant data and meteorological data in one air pollution monitoring station in the city of Athens. The main relationship was between total pollution and high humidity in combination with the low-velocity wind. Through the CCA approach, Larson et al. [14] analyzed an eleven year long measurement time series of waves and profiles from Duck North Carolina in

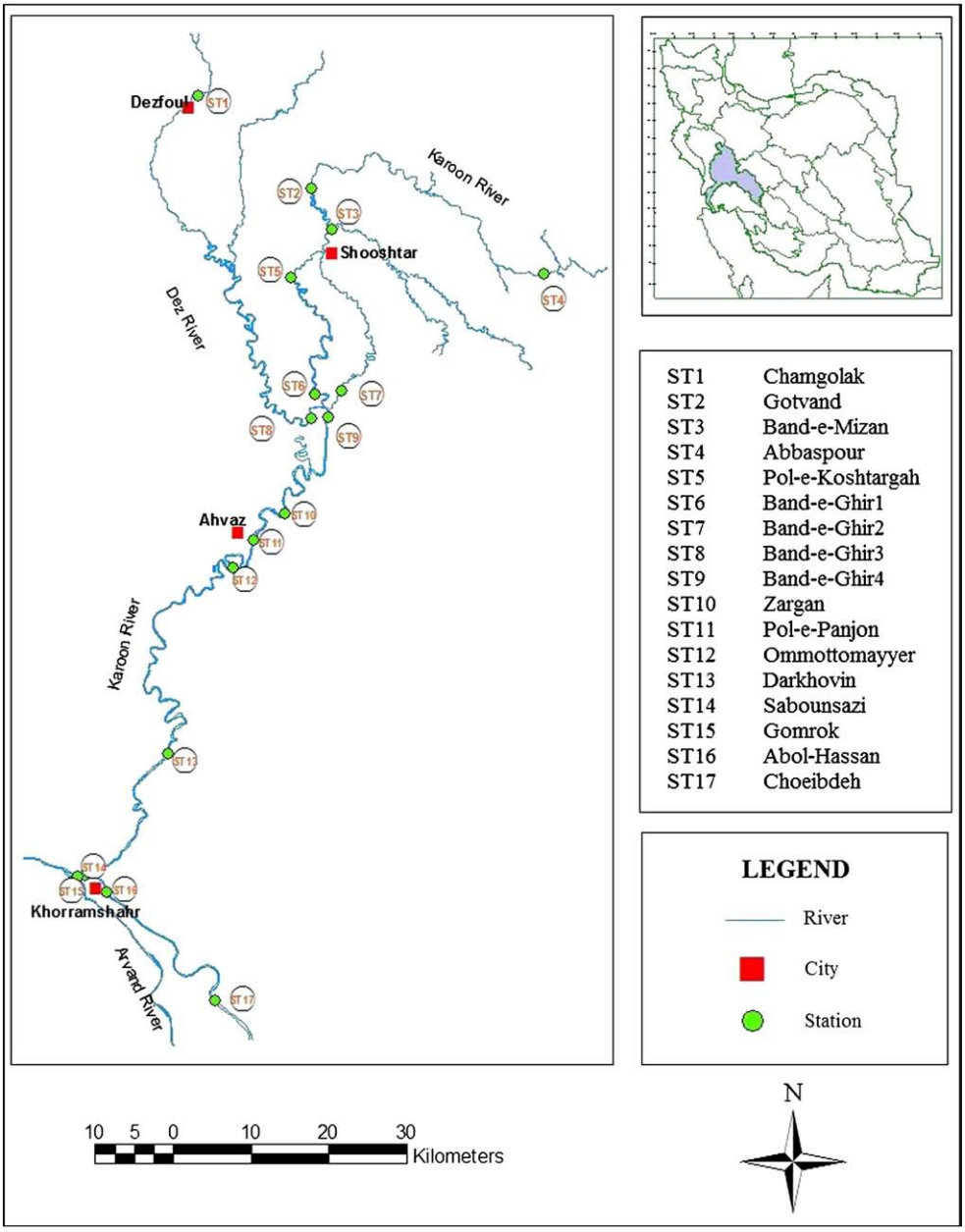

Fig. 1. Territorial lay out of the Karoon River and the location of the river sampling sites.

order to determine covariability between waves and profile response. Liu et al. [15] applied CCA to investigate relationship between personal exposure to 10 volatile organic compounds and biochemical liver tests.

Considering the above-cited studies, the research aims to investigate: (1) identification of the most informative water quality monitoring stations in the Karoon River; (2) determination of the most important water quality parameters in the river; and (3) exploration of relationship between physical and chemical parameters.

2. Methodology

2.1. Case study and data

The Karoon River basin, drained by the Karoon River, is located in the southwest of Iran (Fig. 1). The basin area is 67,000 km^2 and located between longitudes 48°15′ and 52°30′ E, latitude 30°17′ and 33°49′ N [16]. The main branch of the Karoon River originates from the Zagros Mountains and flows down to the northwest of the Persian Gulf. Karoon is the longest river in Iran with an average flow of 90 m^3/s in dry season (Jun–Nov); and 2500 m^3/s in wet season (Dec–May) [17,18]. Karoon supplies drinking water for the public as well as agricultural and industrial activities. Sixty cities and several industries including steel, oil, petrochemical, sugarcane, paper and cement industries, as well as intensely irrigated agricultural areas use Karoon's water and discharge wastewater into it. Construction of several dams along Karoon has increased withdrawal of water from the river. In addition, discharges of domestic, industrial, and agricultural wastewaters have resulted in serious deterioration of water quality, especially downstream of industrial plants in the river [18]. Water quantity and quality of the Karoon River play important roles in sustainable agricultural and industrial development of Khouzestan province especially Ahvaz, Khorramshahr and Abadan cities [17]. Therefore water quality monitoring and finding procedures to limit water pollution must be emphasized. From 1993 to 2002, data (nineteen parameters for water quality assessment) have been collected from 19 water quality monitoring stations along Karoon River. Of these, due to data continuity in measurements, 12 parameters at 17 stations are used in our investigation (Fig. 1). The selected water quality parameters includes water temperature (T), total dissolved solids (TDS), dissolved oxygen (DO), Turb, electrical conductivity (EC), pH, 5 day biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), chloride ion (Cl^-), hardness (Hrad), SO_4^{-2} and NO_3^- ions. Data were not normally distributed and positively skewed. Therefore, annual median values for each parameter are used. The choice of using median rather than mean was based on the fact that the values were quite skewed [19]. Table 1 illustrates basic statistics for a 4-year data set (1999 to 2002) on the Karoon river water quality.

2.2. Principal component analysis

PCA is a multivariate statistical method which can be used for reducing complexity of input variables when there is a large volume of information and it is intended to have a better interpretation of variables [20,21]. In mathematical terms, PCA involves the following five major steps: (1) start by coding the variables X_1, X_2, ..., X_p to have zero means and unit variance; (2) calculate the correlation matrix $\mathbf{R}$; (3) find the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ and the corresponding eigenvectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$ by solving Eq. (1):

$$|R-I\lambda| = 0 \tag{1}$$

(4) discard any components that only account for a small proportion of the variation in data sets; and (5) develop the factor loading matrix and perform a Varimax rotation on the factor loading matrix to infer the principal parameters [10,22]. Details for mastering the art of PCA are published elsewhere [23,24].

2.3. Canonical correlation analysis

In some sets of multivariate data the variables are divided naturally into two groups (i.e. response data and predictor variable). A canonical correlation analysis can then be used to investigate relationships between the two groups. As an exploratory tool, it is used as a data reduction method. The goal of CCA is to construct two new sets of canonical variates $U=\boldsymbol{\alpha}\mathbf{X}$ and $V=\boldsymbol{\beta}\mathbf{Y}$ that are linear combinations of the original variables such that the simple correlation between U and V is maximal, subject to the restriction that each canonical variate U and V has unit variance (to ensure uniqueness, except for sign) and is uncorrelated with other constructed variates within the set. Assume that the $(p+q)\times(p+q)$ correlation matrix between the variables X_1, X_2, ..., X_p and Y_1, Y_2, ..., Y_q takes the following form when it is calculated from the sample for which the variables are recorded:

	X_1 X_2 ... X_p	Y_1 Y_2 ... Y_q
X_1 X_2 . . X_p	p × p matrix A	p × q matrix C
Y_1 Y_2 . . Y_q	q × p matrix C′	q × q matrix B

From this matrix a $q\times q$ matrix $\mathbf{B^{-1}C'A^{-1}C}$ can be calculated, and the eigenvalue problem can be considered as:

$$\left(\mathbf{B}^{-1}\mathbf{C}'\mathbf{A}^{-1}\mathbf{C}-\lambda\mathbf{I}\right)\boldsymbol{b} = 0. \tag{2}$$

It turns out that the eigenvalues $\lambda_1>\lambda_2>\ldots>\lambda_r$ are then the squares of the correlations between the canonical variates. The r subscribe is the smaller of p and q. The corresponding the eigenvectors, $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_r$ give the coefficients of the Y variables for canonical variates. The coefficients of linear combination of X variables (U_i), the ith canonical variate for the X variables, are given by the elements of the $\boldsymbol{a}_i$ vector.

$$\boldsymbol{a}_i = \mathbf{A}^{-1}\mathbf{C}\boldsymbol{b}_i. \tag{3}$$

In these calculations it is assumed that the original X and Y variables are in a standardized form with means of zero and standard deviations of unity. The coefficients of the canonical variates are for these standardized X and Y variables.

Table 1
Water quality parameters for surface water of the Karoon River basin.

Parameters	Stations																
		ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8	ST9	ST10	ST11	ST12	ST13	ST14	ST15	ST16	ST17
BOD_5	Min	0.5	1	0.5	0.5	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	1	0.5
	Max	8	5	7	7	6	8.5	15	16	14	13	16	18	12	14	18	16	12
	Median	2	2	2.5	1.75	2	3.5	4	3	2.75	3	4	3.5	3.5	3	3.8	4	3.2
pH	Min	7.32	7.15	7.41	7.46	7.62	7.39	7.2	1	6.66	6.9	6.7	6.76	7.12	6.63	7.01	6.43	7.05
	Max	8.76	8.91	8.83	8.53	8.96	8.61	8.68	8.8	8.84	8.99	8.95	838	9.55	8.85	9	9.12	8.55
	Median	8.145	8.32	8.12	8	8.105	8.1	8.08	8.06	8.06	8.15	8.135	8.1	8.3	8.15	8.12	8.23	8.07
COD	Min	4	4	4	4	12	8	4	6	4	6	4	12	12	8	10	8	12
	Max	44	132	76	28	96	132	56	205	152	176	73	76	56	201	193	160	420
	Median	12	8	16	12	36	24	24	36	24	28	32	34	28	28	28	32	44
EC	Min	398	165	705	350	585	499	679	102	105	102	107	92	94	114	116	102	746
	Max	2250	9200	2590	5240	1480	2390	3240	4500	3120	2830	2400	2780	2220	4510	3460	3090	8700
	Median	392.5	498	973	503.5	768	1450	1325	1525	1190	1400	1450	1620	1365	1730	1700	1560	2610
TDS	Min	478	612	738	148	351	299	182	158	263	642	186	624	619	684	714	738	448
	Max	1350	1091	1554	383.4	888	1434	1944	2700	1872	1698	1440	1668	2180	2706	2076	2358	5220
	Median	250	397	611	302.1	460.7	882	804	990	714	882	984	984	867	1071	1053	960	1566
Turb	Min	1	10	264	10	10	30	2.71	1.56	1	25	5.33	19	11	21	8	2.07	42
	Max	438	547	402	401	99	999	602	952	259	546	999	578	945	485	899	784	457
	Median	32	51	56	29	49.5	89	98	76	56	105	115	140	160	59	70.5	203	769
DO	Min	1.7	1.2	3.15	4.97	4.44	4.21	3.6	6.19	5.89	0.74	3.95	3.11	5.89	5.98	5.62	5.36	5.42
	Max	10.8	11.55	11.08	9.45	9.96	9.29	9.04	12.14	10.39	9.9	16.6	9.82	11.8	9.53	10.21	9.21	9.69
	Median	8.48	8.2	8.2	8.45	8.3	6.62	6.42	7.76	7.585	7.69	7.3	7.31	7.61	7.31	7.5	7.55	6.31
Cl^-	Min	37	74	101	41	111	175	103	207	118.4	85	142	73	166	146	204	75	380
	Max	383	352	435	152.65	378	404	1140	930	722	750	508	515	915	945	995	1137	11280
	Median	57	159.5	230	101.5	149	337.5	263	391	220	259	275	293	352.5	393.5	426	300	572
T	Min	14.3	14.3	12.9	13.5	13.9	15	14.7	12.7	13.2	13.1	13.1	13.4	14.3	13.7	14.7	14.3	15.4
	Max	26.5	28	26.5	23.7	209.4	28.5	203	29	29	29.7	30.5	30.8	31.7	31.1	30.9	178	33.6
	Median	17.1	18.85	21	17	20.25	23.1	24.55	23.2	22.5	22.75	22.55	22.8	22.85	23.4	24.1	23.1	25.45
Hrad	Min	11	164	24	26	152	36	12	24	1	12	160	12	180	13	3	156	174.1
	Max	492	560	500	235	500	445	900	770	720	564	580	608	792	669.5	678	660	9918
	Median	187	268	239	154	290	290.85	376	377.5	283	325	409	359	466.5	369	356	430	461.7
NO_3^-	Min	1.41	1.0768	1.2266	2.21	1.0768	2.32	0.16	0.3301	1.5458	2.0148	1.2753	1.0588	1.2031	1.167	1.18	2.7363	1.39
	Max	348.66	506.9	585.08	4.98	12.495	5.3	24.521	571.09	530.17	567.02	578.71	574.87	526.18	904.91	509.43	519.51	6.82
	Median	3.865	6.92	4.45	2.98	7.8	3.82	4.22	6.86	6.01	4.645	5.325	4.63	7.6783	4.89	4.065	7.365	3.21
SO_4^{-2}	Min	5.529	8.9702	7.7801	13.87	41.6	54.23	65.11	8.062	7.426	7.298	6.896	6.799	7.587	7.426	7.282	7.699	86.54
	Max	451.44	505.7	465.33	90.7	275.2	425.58	524.72	585.08	902.28	500.21	502.52	522.3	569.45	520	485.13	494.06	1555
	Median	51.2	82.26	87.04	44.15	93.44	149.67	211.26	181.28	122.24	185.6	243.2	192.76	215.35	219.52	210	269.44	269.09

3. Result and discussion

3.1. Identification of important monitoring stations

Calculating the correlation symmetric matrix **R** is the first step in PCA application.

$$\mathbf{R} = \begin{bmatrix}
1.0000 & 0.9942 & 0.9956 & 0.9955 & 0.9945 & 0.9973 & 0.9942 & 0.9731 & 0.9950 & 0.9862 & 0.9969 & 0.9954 & 0.9971 & 0.9979 & 0.9948 & 0.9647 & 0.9934 \\
0.9942 & 1.0000 & 0.9979 & 0.9981 & 0.9993 & 0.9989 & 0.9935 & 0.9571 & 0.9910 & 0.9649 & 0.9952 & 0.9985 & 0.9855 & 0.9981 & 0.9996 & 0.9749 & 0.9933 \\
0.9956 & 0.9979 & 1.0000 & 0.9997 & 0.9982 & 0.9977 & 0.9982 & 0.9630 & 0.9934 & 0.9748 & 0.9983 & 0.9973 & 0.9900 & 0.9990 & 0.9980 & 0.9749 & 0.9975 \\
0.9955 & 0.9981 & 0.9997 & 1.0000 & 0.9976 & 0.9974 & 0.9966 & 0.9588 & 0.9926 & 0.9731 & 0.9976 & 0.9963 & 0.9892 & 0.9990 & 0.9976 & 0.9785 & 0.9974 \\
0.9945 & 0.9993 & 0.9982 & 0.9976 & 1.0000 & 0.9986 & 0.9953 & 0.9607 & 0.9916 & 0.9679 & 0.9965 & 0.9990 & 0.9872 & 0.9984 & 0.9998 & 0.9680 & 0.9935 \\
0.9973 & 0.9989 & 0.9977 & 0.9974 & 0.9986 & 1.0000 & 0.9950 & 0.9677 & 0.9940 & 0.9736 & 0.9962 & 0.9993 & 0.9905 & 0.9986 & 0.9990 & 0.9711 & 0.9935 \\
0.9942 & 0.9935 & 0.9982 & 0.9966 & 0.9953 & 0.9950 & 1.0000 & 0.9739 & 0.9953 & 0.9810 & 0.9981 & 0.9958 & 0.9918 & 0.9960 & 0.9951 & 0.9686 & 0.9974 \\
0.9731 & 0.9571 & 0.9630 & 0.9588 & 0.9607 & 0.9677 & 0.9739 & 1.0000 & 0.9826 & 0.9835 & 0.9709 & 0.9699 & 0.9805 & 0.9622 & 0.9624 & 0.9240 & 0.9671 \\
0.9950 & 0.9910 & 0.9934 & 0.9926 & 0.9916 & 0.9940 & 0.9953 & 0.9826 & 1.0000 & 0.9846 & 0.9970 & 0.9944 & 0.9950 & 0.9926 & 0.9930 & 0.9705 & 0.9963 \\
0.9862 & 0.9649 & 0.9748 & 0.9731 & 0.9679 & 0.9736 & 0.9810 & 0.9835 & 0.9846 & 1.0000 & 0.9826 & 0.9715 & 0.9944 & 0.9767 & 0.9680 & 0.9347 & 0.9791 \\
0.9969 & 0.9952 & 0.9983 & 0.9976 & 0.9965 & 0.9962 & 0.9981 & 0.9709 & 0.9970 & 0.9826 & 1.0000 & 0.9962 & 0.9949 & 0.9978 & 0.9965 & 0.9696 & 0.9985 \\
0.9954 & 0.9985 & 0.9973 & 0.9963 & 0.9990 & 0.9993 & 0.9958 & 0.9699 & 0.9944 & 0.9715 & 0.9962 & 1.0000 & 0.9892 & 0.9974 & 0.9994 & 0.9681 & 0.9932 \\
0.9971 & 0.9855 & 0.9900 & 0.9892 & 0.9872 & 0.9905 & 0.9918 & 0.9805 & 0.9950 & 0.9944 & 0.9949 & 0.9892 & 1.0000 & 0.9921 & 0.9876 & 0.9546 & 0.9912 \\
0.9979 & 0.9981 & 0.9990 & 0.9990 & 0.9984 & 0.9986 & 0.9960 & 0.9622 & 0.9926 & 0.9767 & 0.9978 & 0.9974 & 0.9921 & 1.0000 & 0.9981 & 0.9699 & 0.9952 \\
0.9948 & 0.9996 & 0.9980 & 0.9976 & 0.9998 & 0.9990 & 0.9951 & 0.9624 & 0.9930 & 0.9680 & 0.9965 & 0.9994 & 0.9876 & 0.9981 & 1.0000 & 0.9710 & 0.9941 \\
0.9647 & 0.9749 & 0.9749 & 0.9785 & 0.9680 & 0.9711 & 0.9686 & 0.9240 & 0.9705 & 0.9347 & 0.9696 & 0.9681 & 0.9546 & 0.9699 & 0.9710 & 1.0000 & 0.9796 \\
0.9934 & 0.9933 & 0.9975 & 0.9974 & 0.9935 & 0.9935 & 0.9974 & 0.9671 & 0.9963 & 0.9791 & 0.9985 & 0.9932 & 0.9912 & 0.9952 & 0.9941 & 0.9796 & 1.0000
\end{bmatrix}$$

After solving Eq. (1), 7 eigenvalues were obtained (Table 2). So for each of the eigenvalue, 7 eigenvectors are calculated. Finally, the obtained eigenvectors can be used for constructing 7 principal components (PCs) from input variables. The characteristics of the PCs are presented in Table 3.

In Table 2, eigenvalues, variance proportion, and cumulative variance proportion are shown. Clearly, the first three components accounted approximately 39.81%, 33.44% and 26.41% of the total variance in the data sets, respectively. These three components together accounted for about 99.66% of the total variance and the rest only accounted for about 0.34%. Table 3 shows the eigenvectors, which assess the coefficients for formation of components. It should be noted that for retaining the PCs, a criterion equal to 10^{-8} is used. In the present work, the correlation coefficient considered significant is one that is greater than 0.60 (or >60%). This conservative criterion is selected because of large study area and highly non-linear and dynamic Karoon River system. The stations with rotated factor correlation coefficients less than this value are not considered as principal stations. Table 3 indicated that stations ST5 (Pol-e-Koshtargah), ST6 (Bande-e-Ghir1), ST10 (Zargan), and ST14 (Sabounsazi) have coefficient values less than 0.60 for all of the PCs. These stations are considered less important in explaining the annual variance of the river water quality, and thereby could be the non-principal stations.

For validating the above findings, the water quality data with and without the four non-principal stations are compared. In this study, two cases are developed for comparisons. In the first case, data from the principal stations are used to formulate the following four relationships by regression: Turb vs. EC, pH vs. NO_3^-, T vs. DO, Turb vs. TDS, EC vs. Hard, EC vs. Cl^-, SO_4^{-2} vs. Hard, and COD vs. TDS. In the second case, data from all stations (principal and non-principal stations) are used to formulate the aforementioned eight relationships by regression. These two cases are then compared to determine if the addition of data from the four non-principal stations improved the regression relationships. Comparison of the relationship between these parameters shows the amount of R^2 increase in the case of principal monitoring stations. This increase is different for each parameter. For example the highest increase is related to the Turb vs. TDS (Table 4).

Table 2
Descriptive statistics of created PCs.

PCs	Eigenvalue	Variance proportion	Cumulative variance proportion
PC1	6.767	39.811	39.811
PC2	5.684	33.439	73.251
PC3	4.489	26.410	99.661
PC4	0.036	0.212	99.873
PC5	0.014	0.085	99.958
PC6	0.006	0.037	99.996
PC7	5.0E-4	0.003	99.999
PC8	2.6E-05	1.5E-4	99.999
PC9	5.0E-06	2.9E-05	99.999
PC10	5.0E-06	2.9E-05	99.999
PC11	1.9E-07	1.1E-06	100
PC12	1.3E-15	7.9E-15	100
PC13	1.1E-15	6.9E-15	100
PC14	6.4E-16	3.7E-15	100
PC15	5.8E-16	3.4E-15	100
PC16	4.9E-16	2.9E-15	100
PC17	2.5E-17	1.5E-16	100

Table 3
Eigenvectors obtained through PCA application.

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
ST_1	**0.645**	0.553	0.524	0.060	−0.021	−0.003	−0.004	−3.1E-03	2.4E-04	7.1E-04	5.7E-05
ST_2	0.572	0.596	0.563	0.011	−0.003	0.007	−0.001	−3.7E-04	7.0E-04	−2.5E-04	−8.9E-05
ST_3	0.598	0.592	0.537	0.035	0.044	−0.002	−0.001	4.5E-04	1.3E-04	5.4E-04	4.5E-05
ST_4	0.588	**0.606**	0.533	0.048	0.030	0.001	0.001	1.9E-03	−5.6E-04	−2.3E-04	−7.1E-05
ST_5	0.586	0.571	0.574	0.009	0.019	0.016	−0.001	−2.5E-04	−5.4E-04	4.5E-04	−1.6E-04
ST_6	**0.606**	0.575	0.549	0.009	−0.015	−0.008	−0.002	5.2E-04	1.1E-04	−7.1E-05	3.5E-04
ST_7	**0.635**	0.571	0.515	0.014	0.074	−0.001	0.001	8.5E-05	−1.9E-04	−5.2E-04	−1.6E-05
ST_8	**0.775**	0.472	0.414	−0.074	0.004	−0.004	−0.002	3.0E-04	5.9E-05	1.6E-04	−3.3E-06
ST_9	**0.663**	0.575	0.477	−0.002	0.005	0.046	0.004	1.7E-04	−9.2E-04	−7.4E-04	−1.5E-05
ST_{10}	**0.751**	0.492	0.428	0.104	0.022	−0.002	−0.006	2.5E-04	4.8E-05	1.3E-06	−7.8E-06
ST_{11}	**0.631**	0.572	0.520	0.042	0.038	0.039	−0.004	−3.8E-04	1.6E-03	8.6E-05	1.0E-05
ST_{12}	**0.608**	0.567	0.555	−0.016	0.003	0.005	0.008	−1.1E-03	−5.6E-04	−6.2E-04	9.4E-05
ST_{13}	**0.691**	0.527	0.489	0.075	−0.004	0.023	0.020	4.5E-05	−1.9E-05	−1.9E-05	−6.7E-06
ST_{14}	**0.602**	0.574	0.552	0.054	0.012	−0.002	−0.001	3.1E-03	2.4E-04	7.9E-04	8.3E-05
ST_{15}	0.588	0.580	0.564	0.001	0.007	0.020	0.002	−1.1E-03	−2.2E-04	−3.6E-04	−1.0E-04
ST_{16}	0.517	**0.751**	0.412	0.015	−0.003	−0.001	0.002	−3.3E-04	3.7E-05	−8.8E-05	2.1E-05
ST_{17}	**0.619**	0.611	0.487	0.039	0.054	0.038	−0.001	−5.0E-06	8.3E-05	1.4E-03	−1.4E-05

3.2. Extraction of important parameters

For selecting the principal water quality parameters, a similar approach is used and twelve parameters are selected for this study. Correlation symmetrical matrix $\boldsymbol{\tau}$ is formed.

$$\mathbf{t} = \begin{bmatrix}
1.0000 & -0.0791 & -0.1183 & -0.0415 & 0.1224 & -0.1898 & -0.0656 & 0.1500 & -0.0088 & 0.5758 & 0.1620 & 0.3016 \\
-0.0791 & 1.0000 & 0.7258 & 0.5085 & 0.2672 & 0.5750 & 0.7069 & -0.5572 & 0.6815 & -0.2323 & 0.5501 & 0.5571 \\
-0.1183 & 0.7258 & 1.0000 & 0.9065 & 0.7029 & 0.8191 & 0.9945 & -0.8174 & 0.9518 & -0.1331 & 0.8765 & 0.8134 \\
-0.0415 & 0.5085 & 0.9065 & 1.0000 & 0.8174 & 0.7591 & 0.9064 & -0.8609 & 0.8650 & 0.0468 & 0.8740 & 0.8428 \\
0.1224 & 0.2672 & 0.7029 & 0.8174 & 1.0000 & 0.5629 & 0.7200 & -0.7773 & 0.6285 & 0.0470 & 0.8761 & 0.7940 \\
-0.1898 & 0.5750 & 0.8191 & 0.7591 & 0.5629 & 1.0000 & 0.8117 & -0.5592 & 0.7403 & 0.1905 & 0.7576 & 0.7740 \\
-0.0656 & 0.7069 & 0.9945 & 0.9064 & 0.7200 & 0.8117 & 1.0000 & -0.8044 & 0.9571 & -0.0935 & 0.8958 & 0.8398 \\
0.1500 & -0.5572 & -0.8174 & -0.8609 & -0.7773 & -0.5592 & -0.8044 & 1.0000 & -0.7678 & 0.2965 & -0.7493 & -0.6617 \\
-0.0088 & 0.6815 & 0.9518 & 0.8650 & 0.6285 & 0.7403 & 0.9571 & -0.7678 & 1.0000 & -0.0535 & 0.8093 & 0.8018 \\
0.5758 & -0.2323 & -0.1331 & 0.0468 & 0.0470 & 0.1905 & -0.0935 & 0.2965 & -0.0535 & 1.0000 & 0.1165 & 0.3342 \\
0.1620 & 0.5501 & 0.8765 & 0.8740 & 0.8761 & 0.7576 & 0.8958 & -0.7493 & 0.8093 & 0.1165 & 1.0000 & 0.9445 \\
0.3016 & 0.5571 & 0.8134 & 0.8428 & 0.7940 & 0.7740 & 0.8398 & -0.6617 & 0.8018 & 0.3342 & 0.9445 & 1.00000
\end{bmatrix}$$

In Table 5 eigenvalues and cumulative variance proportions are shown. In the table 93.83% of the total variance is allocated to the five first factors and 6.17% is assigned to the rest of the components. Table 6 shows the coefficients for formation of components. Similar to the previous section, the correlation coefficient considered significant, is selected greater than 0.60 (or >60%). Table 6 indicated that all water quality parameters are considered important. In this table, the most effective variables in PCs formation are shown in bold. We conclude that all water quality parameters used here are important factors to be considered.

3.3. Relationship between physical and chemical parameters in the Karoon River basin

In the preset investigation, at first, CCA is carried out on all complete sets of site data. There are five variables in the response data set i.e. physical parameters including TDS, DO, Turb, EC and T and seven variables in the predictor set i.e. chemical parameters including BOD_5, COD, Cl^-, NO_3^-, SO_4^{-2}, Hard, and pH. Table 7 represents the results of CCA for physical and chemical variables. Correlation coefficient for canonical variates 1, 2, 3 and 4 were 0.993, 0.822, 0.785 and 0.660, respectively. Correlation coefficient for the fifth canonical variate was 0.41, less than 0.45, and therefore it was neglected in conclusion. Only the first canonical correlation was statistically significant ($p<0.0001$). Although the second, third and fourth canonical correlations are large, they are not statistically significant by Chi-square test. The test statistic for canonical variates 2, 3, and 4 is found to be $\chi_2^2 = 26.94$ with 24 degrees of freedom. The test statistics for canonical variates 3 and 4 ($\chi_3^2 = 16.26$, with 15 degrees of freedom) are even less significant. Therefore, there is no real evidence of any relationships between the physical and chemical variables based on canonical variates 2, 3, and 4. It may seem strange that there is no significance in the results although second, third, and fourth canonical correlations are quite high. The rather small sample size may be a reasonable explanation [23]. However, the dominant variable in the first canonical variate for physical variables (U_1) is EC and the dominant variables in the V_1 (chemical parameters) are Cl^-, SO_4^{-2}, and Hard. All other physical and chemical parameters in this canonical variate exhibited no significant correlations. The second canonical variate has high correlations of the response and predictor sets. In this canonical variate the predictor variables are EC, TDS, and T; and the response variables BOD_5, NO_3^- and COD have the high

Table 4
Comparison of the relationship between the parameters in two cases (all monitoring stations and principal monitoring stations).

		Turb-EC	pH-NO_3^-	T-DO	Turb-TDS	EC-Hard	EC-Cl^-	SO_4^{-2}-Hard	COD-TDS
R^2	All monitoring stations	0.527	0.332	0.741	0.500	0.662	0.906	0.892	0.659
	Principal monitoring stations	0.626	0.338	0.781	0.641	0.693	0.911	0.910	0.682

Table 5
Descriptive statistics of created PCs.

PCs	Eigenvalue	Variance proportion	Cumulative variance proportion
PC1	4.079	33.990	33.990
PC2	3.358	27.979	61.969
PC3	1.387	11.562	73.530
PC4	1.274	10.615	84.145
PC5	1.163	9.691	93.835
PC6	0.291	2.427	96.263
PC7	0.280	2.331	98.594
PC8	0.061	0.505	99.099
PC9	0.049	0.409	99.508
PC10	0.042	0.353	99.861
PC11	0.012	0.103	99.964
PC12	0.004	0.036	100.000

Table 6
Coefficients for formation of PCs.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
pH	−0.0444	0.0757	−0.0101	0.2956	**0.9503**	−0.0356	0.0190	−0.0045	0.0028	0.0052	−0.0010	−0.0007
Turb	0.4447	0.1217	**0.8730**	−0.1366	−0.0171	0.0581	−0.0535	0.0074	0.0069	0.0074	−0.0004	−0.0005
EC	**0.7931**	0.4679	0.3377	−0.0747	−0.0923	0.1172	−0.0461	0.0332	0.0389	−0.0312	0.0328	0.0576
T	**0.6844**	0.6298	0.1431	0.0950	−0.0838	0.0225	−0.2072	0.2343	−0.0038	−0.0002	0.0030	0.0015
BOD5	0.3361	**0.9343**	0.0031	0.0108	0.0696	0.0518	−0.0330	−0.0211	−0.0618	−0.0318	−0.0041	−0.0010
COD	**0.6063**	0.3754	0.3087	0.3082	−0.2762	0.4740	0.0023	0.0042	0.0110	0.0101	0.0019	0.0009
TDS	**0.8020**	0.4817	0.3166	−0.0507	−0.0464	0.1037	−0.0232	0.0108	0.0447	−0.0101	0.0794	−0.0034
DO	−0.4879	**−0.6496**	−0.2304	0.2457	0.0987	0.0043	0.4653	−0.0155	0.0057	−0.0006	0.0007	−0.0001
Cl^-	**0.8851**	0.3451	0.2726	−0.0226	0.0163	0.0047	−0.0981	−0.0597	−0.0457	0.0355	−0.0697	−0.0318
NO_3^-	−0.0124	0.0205	−0.1097	**0.9425**	0.3101	0.0404	0.0348	0.0037	0.0028	0.0042	−0.0006	−0.0004
SO_4^{-2}	0.5647	**0.7330**	0.2506	0.0989	0.1133	0.1284	0.0492	−0.0076	0.1924	0.0489	0.0073	0.0025
Hard	0.5632	**0.6297**	0.3050	0.3106	0.2040	0.1269	−0.0100	−0.0004	0.0491	0.1909	−0.0038	−0.0020

Table 7
Canonical correlation analysis of the data set.

Canonical variates		1	2	3	4	5
Canonical correlation		**0.993**	**0.822**	**0.785**	**0.660**	**0.408**
Chi-square		67.865	26.941	16.256	7.168	1.734
Degree of freedom		35	24	15	8	3
Significant level		<0.001	<0.307	<0.365	<0.519	<0.629
Physical parameters	TDS	−0.457	**1.733**	**−3.238**	**−4.978**	**−7.792**
	DO	0.132	−0.312	−1.661	1.146	0.401
	Turb	0.108	−0.605	−0.926	−1.167	0.690
	EC	**−0.720**	**−2.822**	**4.486**	**7.075**	**5.930**
	T	0.230	**1.393**	**−2.212**	−0.371	1.897
Chemical parameters	BOD_5	−0.019	**1.548**	**1.015**	0.062	0.410
	COD	−0.236	−0.761	0.319	0.890	0.295
	Cl^-	**−0.599**	0.328	0.179	0.826	−0.631
	NO_3^-	0.150	**0.934**	−0.657	1.148	−0.263
	SO_4^{-2}	**−0.511**	−0.554	−0.384	**2.257**	**−2.767**
	Hard	**0.310**	−0.298	**−1.124**	**−4.150**	**2.928**
	pH	−0.027	−0.490	0.205	0.044	−0.929

Table 8
Correlation matrix of the data set.

	pH	Turb	EC	T	BOD_5	COD	TDS	DO	Cl^-	NO_3^-	SO_4^{-2}	Hard
pH	1.000											
Turb	0.122	1.000										
EC	−0.190	0.563	1.000									
T	−0.009	0.628	0.740	1.000								
BOD_5	0.576	0.047	0.191	−0.054	1.000							
COD	0.162	0.876	0.758	0.809	0.117	1.000						
TDS	0.302	0.794	0.774	0.802	0.334	0.944	1.000					
DO	−0.079	0.267	0.575	0.681	−0.232	0.550	0.557	1.000				
Cl^-	−0.118	0.703	0.819	0.952	−0.133	0.877	0.813	0.726	1.000			
NO_3^-	−0.042	0.817	0.759	0.865	0.047	0.874	0.843	0.508	0.907	1.000		
SO_4^{-2}	−0.066	0.720	0.812	0.957	−0.094	0.896	0.840	0.707	0.995	0.906	1.000	
Hard	0.150	−0.777	−0.559	−0.768	0.296	−0.749	−0.662	−0.557	−0.817	−0.861	−0.804	1.000

loading. From the third canonical correlation, EC, TDS, and T of physical variables and Hard, BOD_5, and NO_3^- of chemical parameters have high correlation. Fourth canonical correlation, however, shows dominant variables in the U_4 are EC, TDS, Turb, and DO, whilst Hard, SO_4^{-2} and NO_3^- in the V_4 are dominant. Considering the mentioned results, a regular pattern can be seen. EC and TDS are two dominant physical parameters in the all canonical variates. On the other hand, ions and hardness are highly scored from chemical parameters. In addition, although the fifth canonical variate is neglected in the present investigations, the pattern of predictor and response variable correlations is maintained relatively constant. Canonical variate 5 indicates that TDS, EC, and T are associated with Hard, and SO_4^{-2}. Verifying the ability of CCA, simple correlation factor between all parameters are shown in Table 8. The correlation matrix reveals that a relationship exists between EC, TDS, and T, as well as between COD, Cl^-, SO_4^{-2} and Hard but not between NO_3^- and any physical parameters.

4. Conclusion

Surface water quality data for the Karoon River (in southwest Iran) were evaluated for spatial variations and the relationship between physical and chemical parameters. Principal component analysis was applied to determine important monitoring stations and water quality parameters. Canonical correlation analysis was used to find out the relationship between two data sets, i.e. physical and chemical parameters. Results of PCA indicated that ST5 (Pol-e-Koshtargah), ST6 (Bande-e-Ghir1), ST10 (Zargan), and ST14 (Sabounsazi) monitoring stations were non-principal stations and all of water quality parameters were principal. These findings were supported by simple regressions as well. Furthermore, EC and TDS were important physical parameters and ions and hardness were significant chemical parameters in canonical variates. To verify the ability of CCA simple correlation was also used. The methods used here can offer an effective solution to water quality management in cases where there is large complex water quality data involved.

References

[1] R. Noori, M.A. Abdoli, A. Ameri, M. Jalili-Ghazizade, Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: a case study of Mashhad, Environmental Progress & Sustainable Energy 28 (2009) 249–258.
[2] R. Noori, A.R. Karbassi, A. Farokhnia, M. Dehghani, Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques, Environmental Engineering Science 26 (2009) 1503–1510.
[3] O.P. Bricker, B.F. Jones, Main factors affecting the composition of natural waters, in: B. Salbu, E. Steinnes (Eds.), Trace Elements in Natural Waters, CRC Press, Boca Raton, FL, 1995, pp. 1–5.
[4] S. Shrestha, F. Kazama, Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan, Environmental Modelling & Software 22 (2007) 464–475.
[5] W. Dixon, B. Chiswell, Review of aquatic monitoring program design, Water Research 30 (1996) 1935–1948.
[6] A. Facchinelli, E. Sacchi, L. Mallen, Multivariate statistical and GIS-based approach to identify heavy metals sources in soils, Environmental Pollution 114 (2001) 313–324.
[7] V. Simeonov, J.A. Stratis, C. Samara, G. Zachariadis, D. Voutsa, A. Anthemidis, M. Sofoniou, T. Kouimtzis, Assessment of the surface water quality in northern Greece, Water Research 37 (2003) 4119–4124.
[8] S. Gangopadhyay, A.D. Gupta, M.H. Nachabe, Evaluation of ground water monitoring network by principal component analysis, Ground Water 39 (2001) 181–191.
[9] M. Terrado, D. Barcelo, R. Tauler, Identification and distribution of contamination sources in the Ebro river basin by chemometrics modelling coupled to geographical information systems, Talanta 70 (2006) 691–704.
[10] Y. Ouyang, Evaluation of river water quality monitoring stations by principal component analysis, Water Research 39 (2005) 2621–2635.
[11] H. Hotelling, Relation between two sets of variates, Biometrica 28 (1936) 321–329.
[12] H.R. Glahn, Canonical correlation analysis and its relationship to discriminant analysis and multiple regression, Journal of Atmospheric Sciences 25 (1968) 23–31.
[13] M. Statheropoulos, N. Vassiliadis, A. Pappa, Principal component and canonical correlation analysis for examining air pollution and meteorological data, Atmospheric Environment 32 (1998) 1087–1095.
[14] M. Larson, M. Capobianco, H. Hanson, Relationship between beach profiles and waves at Duck, North Carolina, determined by canonical correlation analysis, Journal of Marine Geology 163 (1999) 275–288.
[15] J. Liu, W. Drane, W. Liu, T. Wu, Examination of the relationships between environmental exposures to volatile organic compounds and biochemical liver tests: application of canonical correlation analysis, Journal of Environmental Research 109 (2009) 193–199.
[16] K. Naddafi, H. Honari, M. Ahmadi, Water quality trend analysis for the Karoon River in Iran, Environmental Monitoring and Assessment 134 (2007) 305–312.
[17] V. Diagomanolin, M. Farhang, M. Ghazi-Khansari, N. Jafarzadeh, Heavy metals (Ni, Cr, Cu) in the Karoon waterway river, Iran, Toxicology Letters 151 (2004) 63–68.
[18] S.A. Mojahedi, J. Attari, A Comparative Study of Water Quality Indices for Karun River; World Environmental and Water Resources Congress: Great Rivers, Kansas City, Missouri, 2009.
[19] T.W. Anderson, S.L. Sclove, The Statistical Analysis of Data, The Scientific Press, 1986.
[20] R. Noori, A. Khakpour, B. Omidvar, A. Farokhnia, Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic, Expert Systems with Applications (2010) DOI:10.1016/j.eswa.2010.02.020.
[21] R. Noori, M.A. Abdoli, M. Jalili-Ghazizade, R. Samifard, Comparison of ANN and PCA based multivariate linear regression applied to predict the weekly municipal solid waste generation in Tehran, Iranian Journal of Public Health 38 (2009) 74–84.
[22] R. Noori, A.R. Karbassi, M.S. Sabahi, Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste predicting, Journal of Environmental Management 91 (2010) 767–771.
[23] B.F.J. Manly, Multivariate Statistical Methods: A Primer, Chapman & Hall, London, 1986.
[24] B.G. Tabachnick, L.S. Fidell, Using Multivariate Statistics, Allyn and Bacon, Boston, London, 2001.